(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 April 2001 (19.04.2001)

PCT

(10) International Publication Number
WO 01/27809 A2

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/NL00/00742

(22) International Filing Date: 16 October 2000 (16.10.2000)

(25) Filing Language: Dutch

(26) Publication Language: English

(30) Priority Data:
1013297 15 October 1999 (15.10.1999) NL

(71) Applicant *(for all designated States except US)*: **PLANT RESEARCH INTERNATIONAL B.V.** [NL/NL]; Droevendaalsesteeg 1, NL-6708 PB Wageningen (NL).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **VAN DER KRIEKEN, Wilhelmus, Maria** [NL/NL]; Laantje van Anton Pieck 2, NL-6708 RE Wageningen (NL). **KODDE, Jan** [NL/NL]; Van Doesburglaan 28, NL-6708 MC Wageningen (NL).

(74) Agents: LAND, Addick, Adrianus, Gosling et al.; Arnold & Siedsma, Sweelinckplein 1, NL-2517 GK Den Haag (NL).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/27809 A2

(54) Title: VISUALIZING RELATIONS IN DATA SETS

(57) Abstract: The present invention relates to a method for comparing and/or analyzing data files, such as for instance are obtained in one or more array forms or in a matrix form, wherein: the data and relations from the arrays or matrices with mutual correlations are placed via rearrangement in a virtual matrix, whereby these mutual relations or correlations can be made easily visible for a user, for instance on a screen.

VISUALIZING RELATIONS IN DATA SETS

Introduction

Comparison of data files shows whether there are similarities or differences in the files. Generally the mutual relation between two sets of data is compared, and sometimes between a number of sets. This comparison is often a time-consuming activity ("data-mining") wherein errors can easily be made. The invention described here relates to a technique for comparing two or more data sets by showing the differences and/or similarities between the sets in figures. This technique is for instance important for comparing data from experiments with micro-arrays, although other data sets can also be compared with this technique. Other examples are data sets from:

cosmology,
mathematics,
screening of the population,
patient screening,
sociological research,
physical determinations,
biotechnology,
High throughput systems.

Analysis of data files of DNA micro-arrays (and DNA chips) are further elaborated here. Files of DNA-micro-arrays are an example of biotechnological data files (other examples hereof are for instance files of dot blots, cDNA-AFLP, Northern blots, Southern blots and protein-arrays from proteomics). The use of DNA micro-arrays is growing enormously in molecular genetics. In this technique DNA fragments associated with particular genes are placed as little "spots" on a glass plate (the sequence of the DNA fragments and function of the associated gene can be known after analysis). With particular equipment it is possible to make many copies of such a glass plate (the position of the DNA spots on the glass plate is the same in all copies.) Many

thousands of DNA (or cDNA or RNA) spots can be placed on a small glass plate (of for instance one square inch). The DNA, cDNA or RNA in the spots on the array are further designated as "targets". With micro-arrays it is possible to show via determined standard DNA-techniques (hybridization techniques with labeled probes) which genes are active in a determined tissue, under particular conditions. In brief, this proceeds as follows: the DNA targets on the array are hybridized (compounded) with DNA, RNA or cDNA originating from the tissue for testing. This DNA, RNA or cDNA is marked with a particular colour (for instance yellow) and is therefore referred to as probe. After hybridization with the probe the intensity of the colouring is proportional to the (relative) expression level of the gene. Two DNA or RNA probes can also be tested simultaneously. The second probe originates for instance from tissue which has undergone a treatment other than the tissue from which the first probe is made. The second probe can be marked with a different colour (for instance red). Targets on the micro-array which are yellow or red after the hybridization are only expressed in one of the tissues and targets with an intermediate colour (orange-like colours) are expressed in both tissues. With optical equipment the colour of a determined substance can be measured very precisely and the mutual expression difference in the treated and untreated tissue can hereby be determined per target. The differences in expression can be easily recorded: genes which are expressed specifically enough under particular conditions (for instance from tissue after administering of a determined medicine) can be isolated and then studied further. There are different variants for using DNA micro-arrays (cDNA micro-arrays, subtractive cDNA micro-arrays, micro-arrays wherein EST, expressed sequence tags, are spotted on the glass plate as targets etc.). Because large numbers of genes are compared (often many thousands per array and these numbers are becoming ever greater) large data files

are created. These data files have heretofore been processed in rather unsophisticated manner. Per target the change in expression in different conditions is analyzed and stored (via a spreadsheet, via a coloured representation of the array, wherein the colour represents the change in gene expression or representation in a "tree diagram"). Because this method of comparison is very time-consuming (the many thousands of targets must be mutually compared) and because errors can easily occur, in practice a pre-formulated expression profile is often sought and the strongest exponent hereof is selected, resulting in the loss of much information. In such a one-dimensional analysis the mutual relations between the different expression patterns therefore often do not come to light very well, while this is precisely the valuable information which can be generated with micro-arrays. Even if only one specific expression profile is being sought, it is very useful to know whether this profile can be placed in a larger context. It may for instance be important to know whether genes with a specific expression pattern are involved in a plurality of processes. This also reduces the chance of the found profile being the result of measurement errors.

Object of the data analysis program specified here

The object of the analysis program described here is to visualize expression data obtained using micro-arrays (or data from other data files) in the form of virtual arrays. The array elements are herein rearranged in a manner such that genes with a comparable expression pattern are grouped together at a meaningful position in the virtual array. Mutual relations hereby become visible in rapid, ordered and clear manner.

This is achieved by rearranging the targets of the original array on a virtual array. The targets associated with a treatment with a determined specificity (for instance specific gene expression after administering of a medicine) are herein placed together

in a figure representing a virtual array. The targets which are expressed in another treatment are also placed together, and overlaps within these groups of genes are also visualized. Relations in gene expression in different treatments hereby become immediately apparent. Genes which are always expressed together and form, as it were, networks of gene expression are also identified. The rearrangement is such that the information about the original position of the targets is saved so that it is possible to retrieve which gene (function and sequence) is specifically expressed. The new location of the target in this virtual array is determined by the relative expression level relative to the expression of the control genes or of genes which are expressed in other treatments. In a single (1 determination/1 physiological condition) experiment the results are easy to compare. If however a plurality of mutually related conditions (for instance a time array and/or a concentration array) have to be compared, the location of the target in the virtual array also depends on the result of the other measurements within the inputted array.

Further advantages, features and details of the present invention will be elucidated on the basis of the following description with reference to the annexed drawings, in which:

fig. 1 shows a diagram of a regeneration process in a culture medium as embodiment of the present invention,

fig. 2 shows a graphic representation of the result of the experiment according to fig. 1,

fig. 3 is a graphic representation of regrouping of the representation of fig. 2,

fig. 4 is a graphic representation of the groups of genes derived from fig. 3 which are expressed,

fig. 5 is a graphic representation of a regrouping obtained according to method 1 below, and

fig. 6 is a graphic representation of a regrouping obtained according to method 2 below.

The foregoing is illustrated with an experiment. Administering of the plant hormones auxin and cytokinin to plant tissue in vitro results in regeneration of roots, shoots and/or white or green callus. These regeneration processes are dependent on the concentration of auxin and cytokinin in the culture medium (Figure 1). Gene expression values associated with an array tested under the conditions as shown in figure 1 are shown graphically in figure 2. The colour indicates, between two threshold values, the height of the expression: (blue is low and red is high expression). The groups of genes which are expressed in relation to a determined treatment are shown grouped together in figure 3. The function for the genes which can be derived herefrom is shown in figure 4. The set-up of the physiological experiment in combination with the rearrangement of the targets on the virtual arrays determines the information which can be obtained. What these figures clearly illustrate is that rearrangement of the results in this manner gives a much clearer picture of possible relations between the genes. Analysis without the technique described here will take up very much time and it is doubtful whether the relations described here are found, since this will depend on the time available and the experience of the researcher performing the analyses. In this experiment only genes are shown whose expression increases relative to gene expression in the untreated tissue. In similar manner, genes which are expressed less relative to this tissue can also be processed in the virtual array. This rearrangement can be easily automated by means of a computer program.

Methods

Shown in figure 1 is a typical example of an experiment wherein gene expression, by means of micro-arrays, is studied in relation to different physiological conditions. Within this data-analysis method virtual arrays are made by ordering individual targets on the arrays such that they correspond as much as possible with

the physiological incubation conditions in the experiment. Hereby the position of a particular target on the virtual array is hereby no longer random but on the contrary correlated to its function at a determined physiological condition in the experiment. Fig. 3 shows the result of an ideal virtual arrangement of the results of fig. 1.

In order to obtain a good arrangement two algorithms for the rearrangement are shown below. A method wherein clusters are generated with targets which have mutually comparable gene expression patterns in tissues incubated under determined conditions. In the second method the targets are rearranged on the basis of their individual expression properties under such conditions.

The algorithms are capable of (virtually) ordering the results obtained in experiments with micro-arrays. The micro-arrays consist of DNA targets having therein copies of complete genes or parts thereof. In the experiments the expression is measured of the genes which are present on the micro-array. In order to measure this expression a known mRNA pool, isolated from tissue, is converted into labeled cDNA. The labeled cDNAs can bind to complementary DNA in the targets. The expression level of a gene determines its proportion in the mRNA pool. This expression level can thus be measured by determining the quantity of label (of the cDNA) on a particular target.

Taken as starting point in the experiments are tissues incubated under different conditions. These incubations are such that they have a determined mutual relation (for instance a concentration array of a determined substance and/or a time array). The example shows an experiment with two substances (auxin and cytokinin). The function and relation of the targets can be determined from the obtained virtual arrangement.

The data are expression levels which are measured on the micro-arrays (quantity of one or more

labels on a target) in the above stated experiments. For the algorithm this data is placed in columns (or rows). Each column (or row) herein represents the data obtained at a determined incubation condition (determined concentration of auxin or cytokinin). A row (or column) represents the expression results of a target under all incubation conditions.

Method 1. Clustering of genes with comparable expression pattern on virtual arrays.

Data transformation

When the data is sorted, the expression values are first classified in a limited number of levels. Each level therefore represents an expression value between two preselected limit values. In those targets where the expression value lies below the lowest limit value, the level value zero is assigned (the value zero is assigned to low expression levels). The final outcome can be optimized by choosing the limit values of the different levels. If down-regulation of genes is also being studied, the zero value can for instance be assigned to another level.

Sorting of the data on the basis of clusters.

The targets in the different columns are clustered. This can be done for instance as follows: A binary code is assigned to each target. One binary number is assigned (0 or 1) to each expression value of a determined column. One of the expression levels, for instance expression level 0, is given the binary number 0 and all other expression levels the value 1. A row with numbers thus results which forms the binary code. All targets with the same binary code are clustered.

Placing of arrays during virtual arrangement.

Prior to the virtual arrangement, the arrays representing the different incubation conditions are placed in a determined sequence/position. This placing represents for instance the test set-up. In this case the incubation with the lowest concentration of substance 1 is placed on the left and that with the highest on the

right (and at the bottom the lowest concentration of substance 2 and at the top the highest concentration of substance 2 (as is usual in the plotting of figures).

Calculation of the optimal position of the clusters on the virtual arrays.

The virtual arrangement of clusters within the arrays can be based on the following. All targets within a cluster have the same, calculated position. If a cluster is expressed at a high concentration of substance 1 and of substance 2, this cluster is then also placed on the right at the top within all virtual arrays. In other words, the "X and Y" coordinates of the cluster (and the targets within this cluster) are dependent on the expression pattern of the clustered targets in the different arrays.

Placing of the clustered targets on the virtual arrays.

Once the X and Y coordinates have been calculated targets can be placed on the virtual arrays. This takes place cluster by cluster (all targets within a cluster have in the first instance the same X and Y coordinates). The sequence in which clusters are placed can be in order of increasing or decreasing X and/or Y value, in accordance with increasing cluster size, and so on. If during placing the calculated position is already occupied, the closest empty position is chosen. The visualizing will be clearer by beginning with the target which has the highest total expression within a cluster.

If during placing of a new cluster the calculated position is occupied, another position is chosen. This position can for instance be the closest position which is still available, although other placing strategies can also be envisaged. All targets within the cluster then acquire the X and Y coordinates of this new position.

For a good overview of the virtual arrays (good delimitation of different clusters), it can be advantageous to make the virtual arrays larger than the

original arrays. In this manner an equal number of virtual and original arrays is generated. The expression values on the virtual arrays are herein visualized by means of colours, grey tones or hatchings which represent the expression level. In addition, the virtual target also contains information about the original target (specification of the DNA, location on the original arrays etc).

The clustering can be based on more than two dimensions, the visualizing (placing of the targets and clusters) is two-dimensional and the expression level can be shown as third dimension (via colour or hatching etc).

Method 2. Placing of individual targets in virtual arrays so that relations in gene expression become apparent in relation to the set-up of the performed experiments.

Sorting and placing of individual targets on the virtual array.

This form of arrangement is based on the calculation of an individual X and Y coordinate for each target in the virtual array. This calculation can for instance be done as follows: of each target the expression on a determined array is multiplied by the position of this array relative to the other arrays in the physiological experiment in the X or respectively Y direction (in two-dimensional analysis, in the case of more dimensions calculations are of course performed with proportionally more coordinates). The expression of a determined target of figure 1 which is placed in the physiological set-up on the left at the bottom is for instance multiplied by 1 to calculate both the x and the y coordinate. The expression of this target in the array to the right thereof is multiplied by 2 for the x coordinate and by 1 for the y coordinate. The expression in the physiological condition hereabove is multiplied by 2 for both the x and the y coordinate. This is calculated for the target for all incubation conditions in the physiological experiment. All outcomes for a determined

target are then added together. Finally, this number is divided by the sum of all expression levels of the target in all incubation conditions. This calculation is performed in order to determine the x and the y position of each target on the virtual array.

Form of the virtual array.

The form of the original array depends on the application onto the glass plate. The form of this array (number of rows and columns) may hereby not match the physiological incubation conditions very well. It may therefore be advantageous to adapt the form of the virtual array to the performed experiments (see for instance figure 1).

Sorting of the data

The targets are sorted on the basis of these X and Y coordinates. The form (number of columns with targets or number of rows with targets) of the virtual array is taken into account here. If for instance the number of columns (equals the width of the array) is taken into account, the targets are first sorted on the basis of increasing or decreasing y coordinates. The sorted targets are then subdivided into groups the size of the array width. Sorting within these groups subsequently takes place on the basis of increasing or decreasing y coordinate.

Placing of targets on the virtual array

The sorted targets are placed in the virtual array row by row or column by column. The starting position (which of the four corners) and manner of placing (rows or columns) depends on the sequence (the x or y coordinate first) and method (increasing or decreasing) of sorting. If a start is made with the x coordinate, the rearrangement is then particularly optimized, with the algorithm used here, for the variable which is plotted on the x-axis. If a start is made with the y coordinate, rearrangement is then particularly optimized for the variable which is plotted on the y-axis. With other sorting and placing methods this may be

different again. In all cases a skilled researcher can analyse the data quickly and in orderly and reliable manner.

During placing the same number of virtual arrays is in principle generated as there are incubation conditions (real arrays) in the experiment. In the virtual array only the position of a target is changed. The change in the position is the same over all arrays (If for instance target with number 321 is placed in the top left-hand corner in array 1, this will also occur in all other arrays). The expression does not change but can be represented in a different manner. It is even recommended to represent the expression by means of colours (wherein a determined colour represents a determined expression range).

Interpolations or extrapolations for making virtual arrays.

It is the case for both method 1 and 2 that it is possible via interpolation or extrapolation to make an array which can represent a determined, non-tested physiological condition. The expression data of a determined target are herein calculated via interpolation or extrapolation of expression levels which have been measured (associated with incubation conditions which are related to those of the calculated array).

CLAIMS

1. Method for comparing and/or analyzing data files, such as for instance are obtained in one or more array forms or in a matrix form, wherein:

- the data and relations from the arrays or matrices with mutual correlations are placed via rearrangement in a virtual matrix, whereby these mutual relations or correlations can be made easily visible for a user, for instance on a screen.

2. Method as claimed in claim 1, wherein the data files originate from DNA micro-arrays or DNA chips.

3. Method as claimed in claim 2, wherein expression values originating from the DNA micro-arrays are classified in a limited number of levels, whereby a binary code is preferably obtained.

4. Method as claimed in claim 3, wherein targets are clustered in different columns of the micro-array.

5. Method as claimed in any of the foregoing claims, wherein prior to the virtual arrangement the arrays are placed in a determined sequence in accordance with different incubation conditions.

6. Method as claimed in claim 2, wherein at least one individual x and y coordinate is calculated for each target in the virtual array.

7. Method as claimed in claim 6, wherein the expression of a determined array is multiplied by the position of this array relative to the preceding arrays.

8. Method as claimed in claim 7, wherein targets are sorted on the basis of a first (x) respectively a second (Y) coordinate and the sorted targets are placed row by row, column by column, in the virtual array.

9. Computer system, in which is stored a program for performing one or more of the above methods.

10. Data file obtained according to the method from one or more of the above claims.

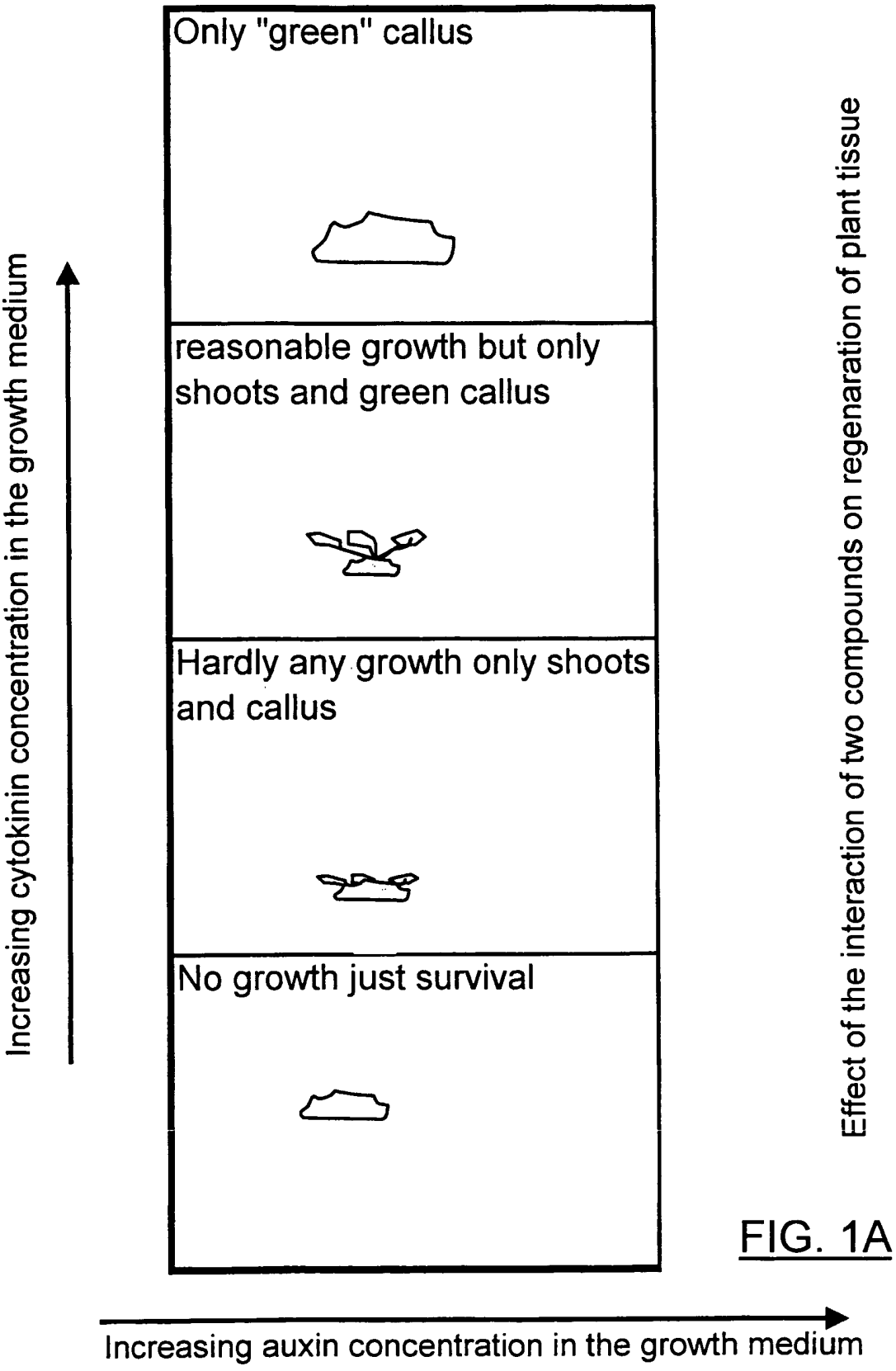

Only "green" callus
reasonable growth but only shoots and green callus
Hardly any growth only shoots and callus
No growth just survival
Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium
Effect of the interaction of two compounds on regenaration of plant tissue

FIG. 1A

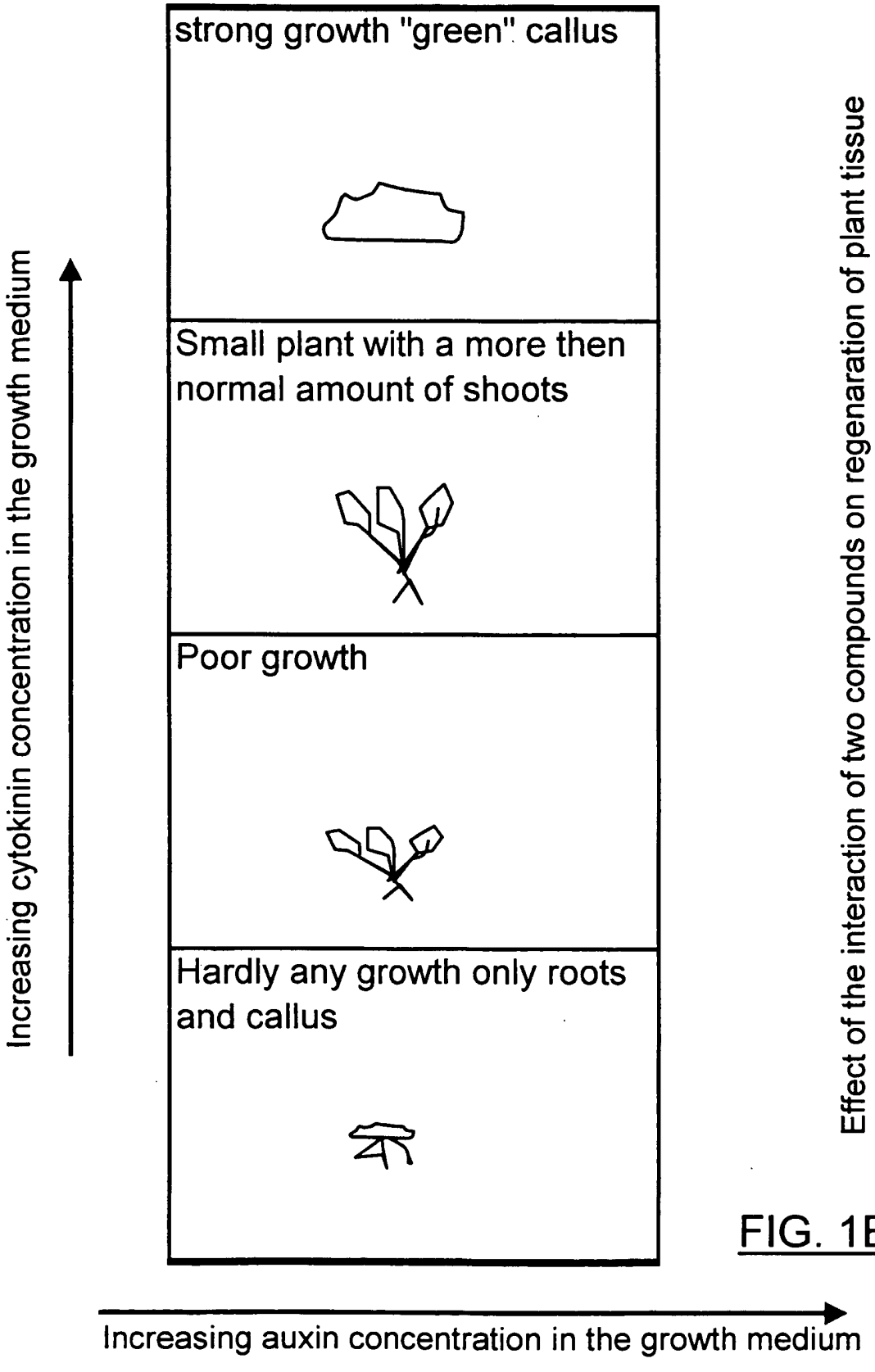

strong growth "green" callus
Small plant with a more then normal amount of shoots
Poor growth
Hardly any growth only roots and callus
Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium
Effect of the interaction of two compounds on regenaration of plant tissue

FIG. 1B

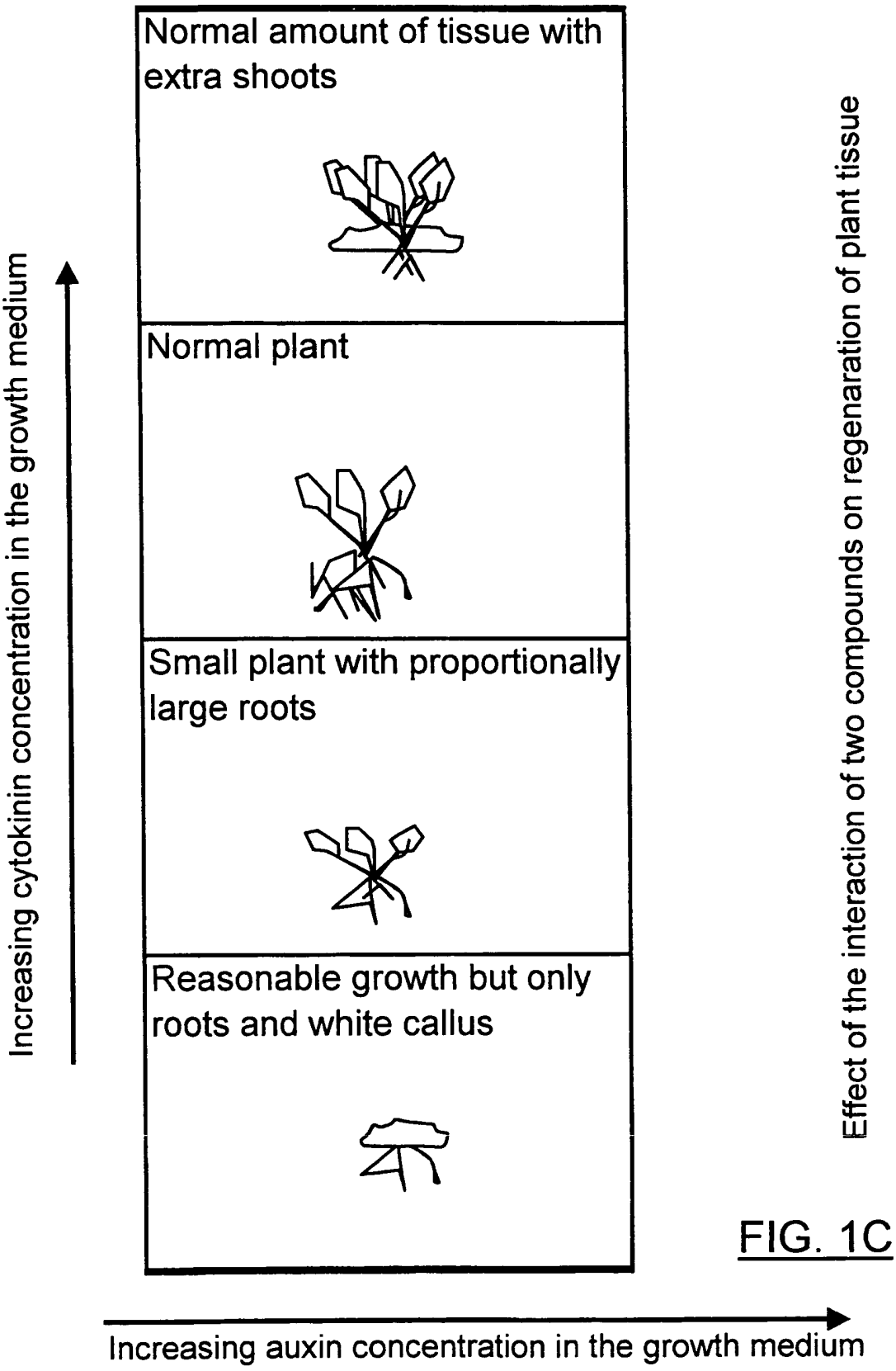

Normal amount of tissue with extra shoots
Normal plant
Small plant with proportionally large roots
Reasonable growth but only roots and white callus
Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium
Effect of the interaction of two compounds on regenaration of plant tissue

FIG. 1C

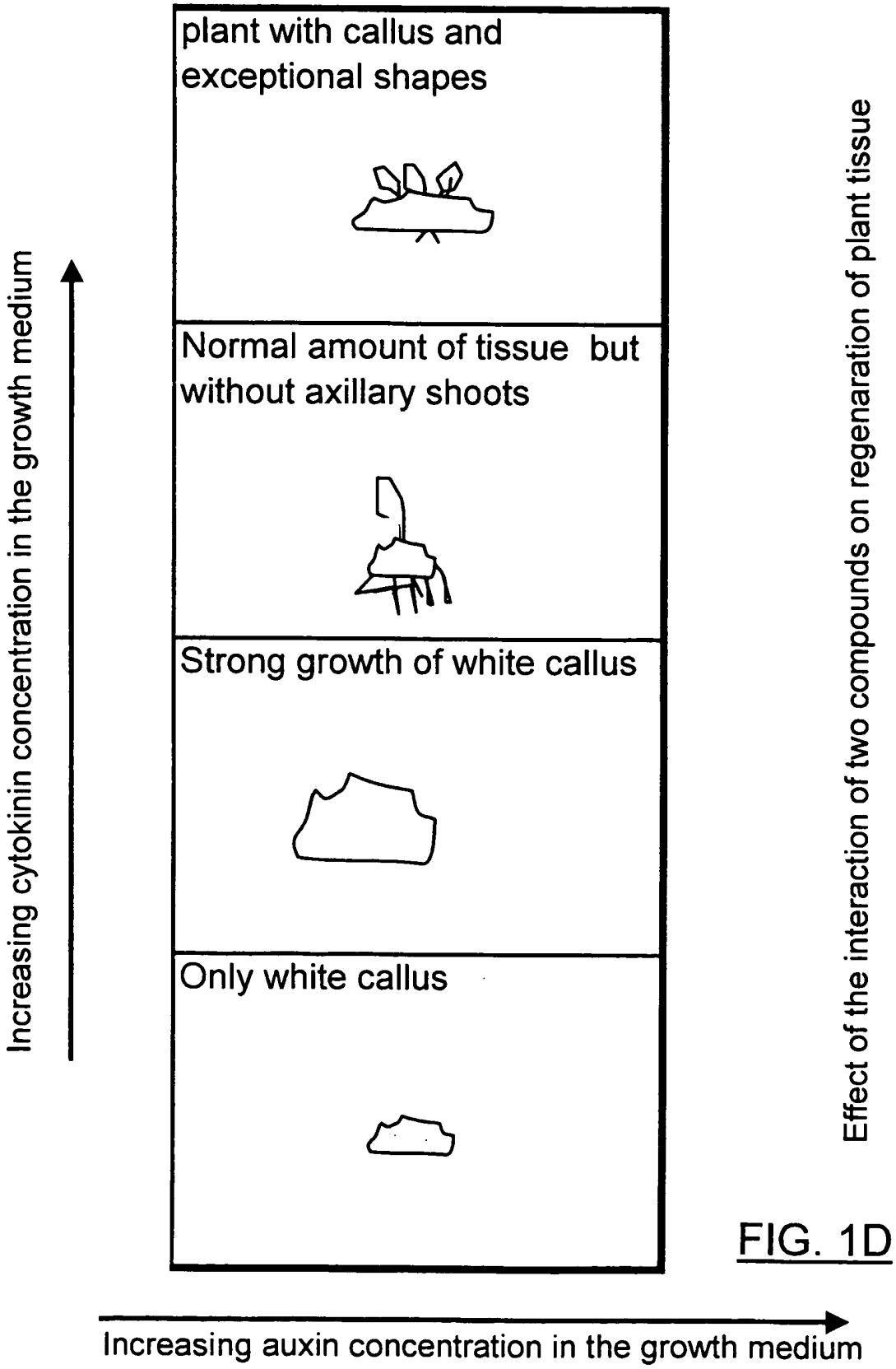

plant with callus and exceptional shapes
Normal amount of tissue but without axillary shoots
Strong growth of white callus
Only white callus
Increasing cytokinin concentration in the growth medium
Effect of the interaction of two compounds on regenaration of plant tissue
Increasing auxin concentration in the growth medium

FIG. 1D

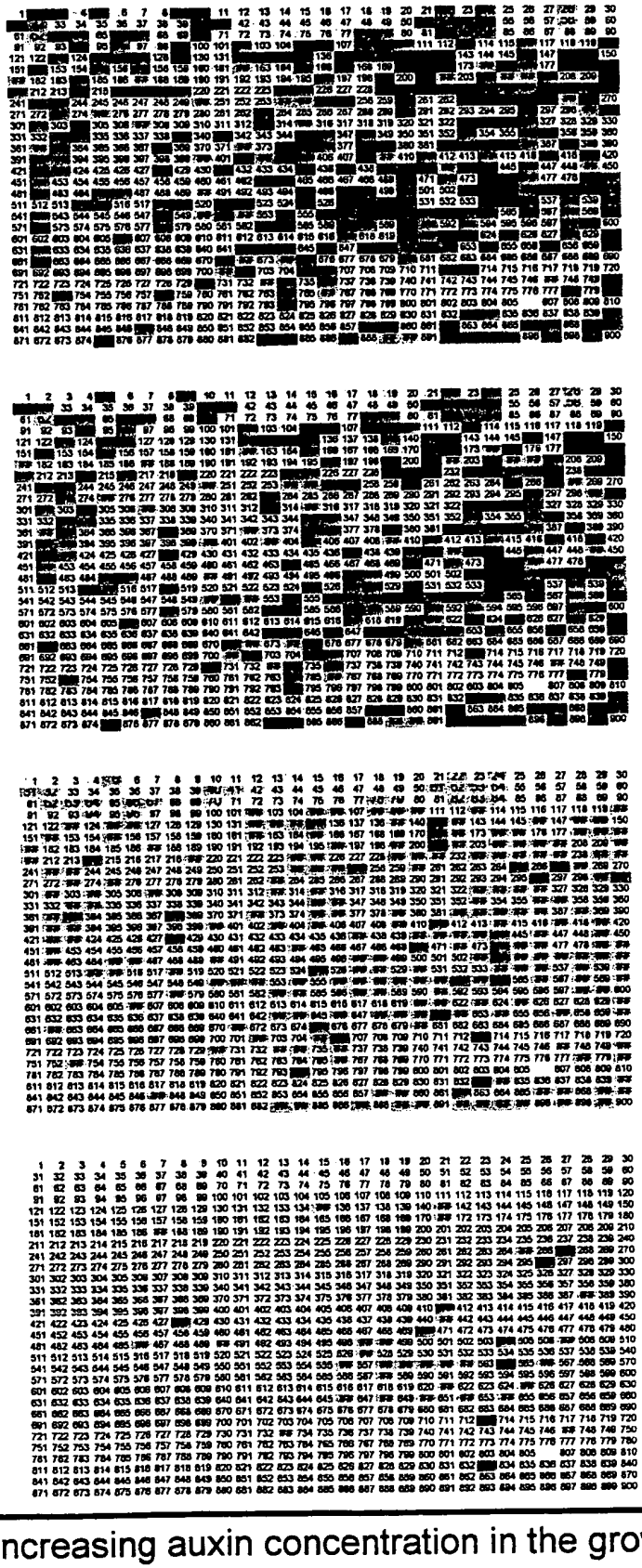

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium
Microarray data before rearrangement.

FIG. 2A

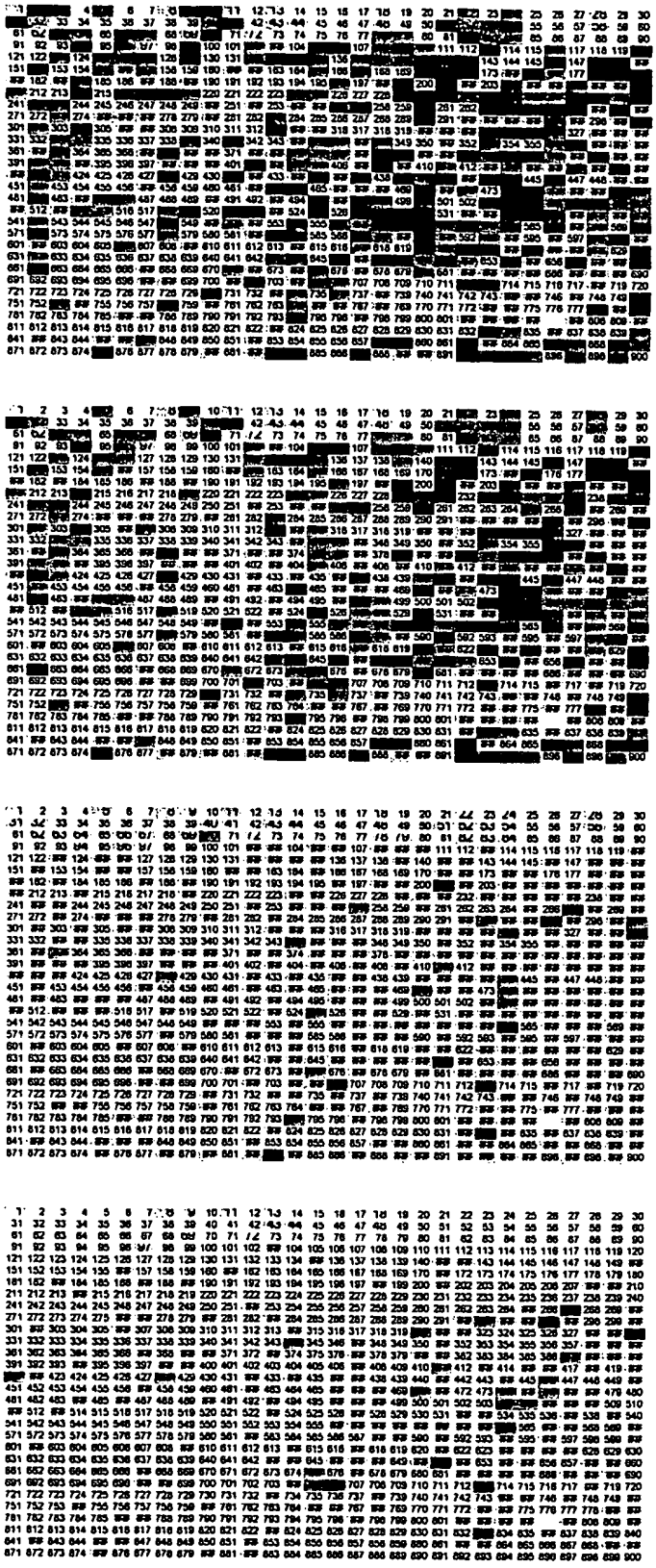

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium
Microarray data before rearrangement.

FIG. 2B

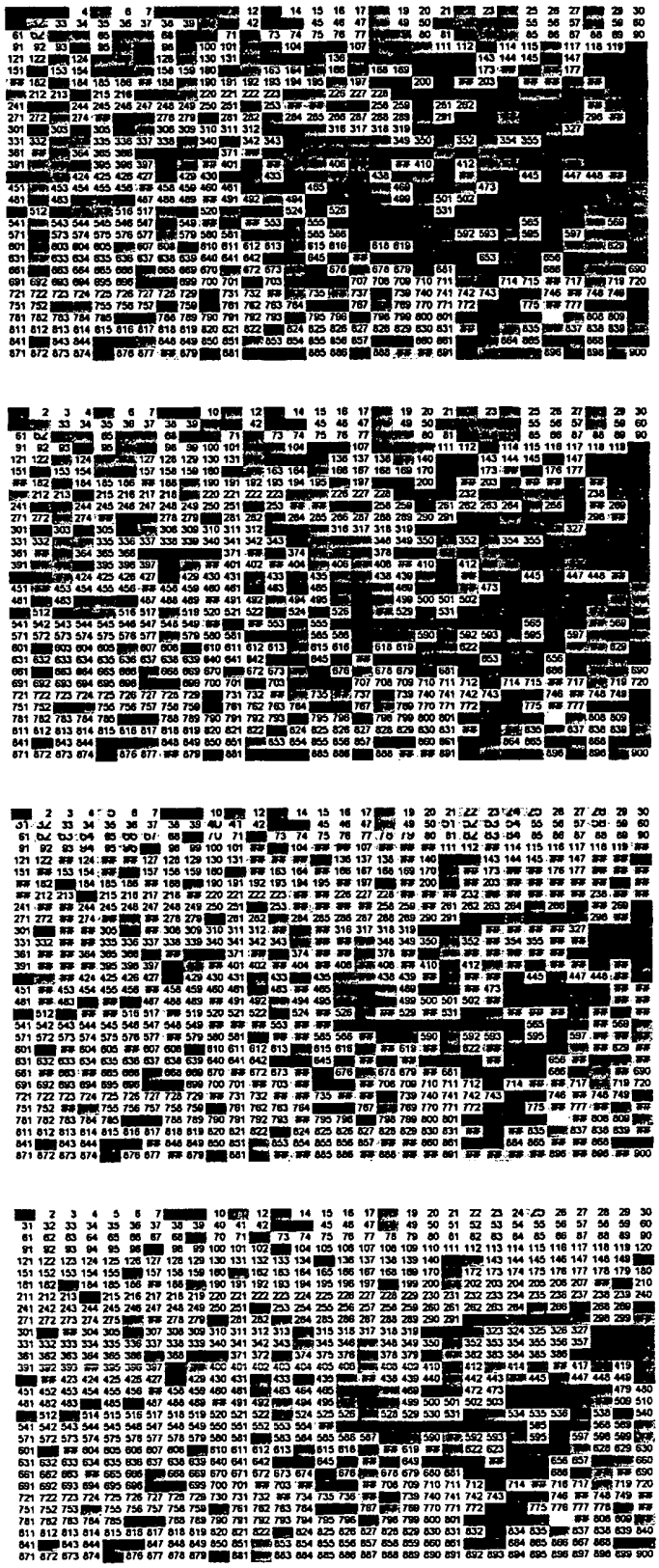

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium
Microarray data before rearrangement.

FIG. 2C

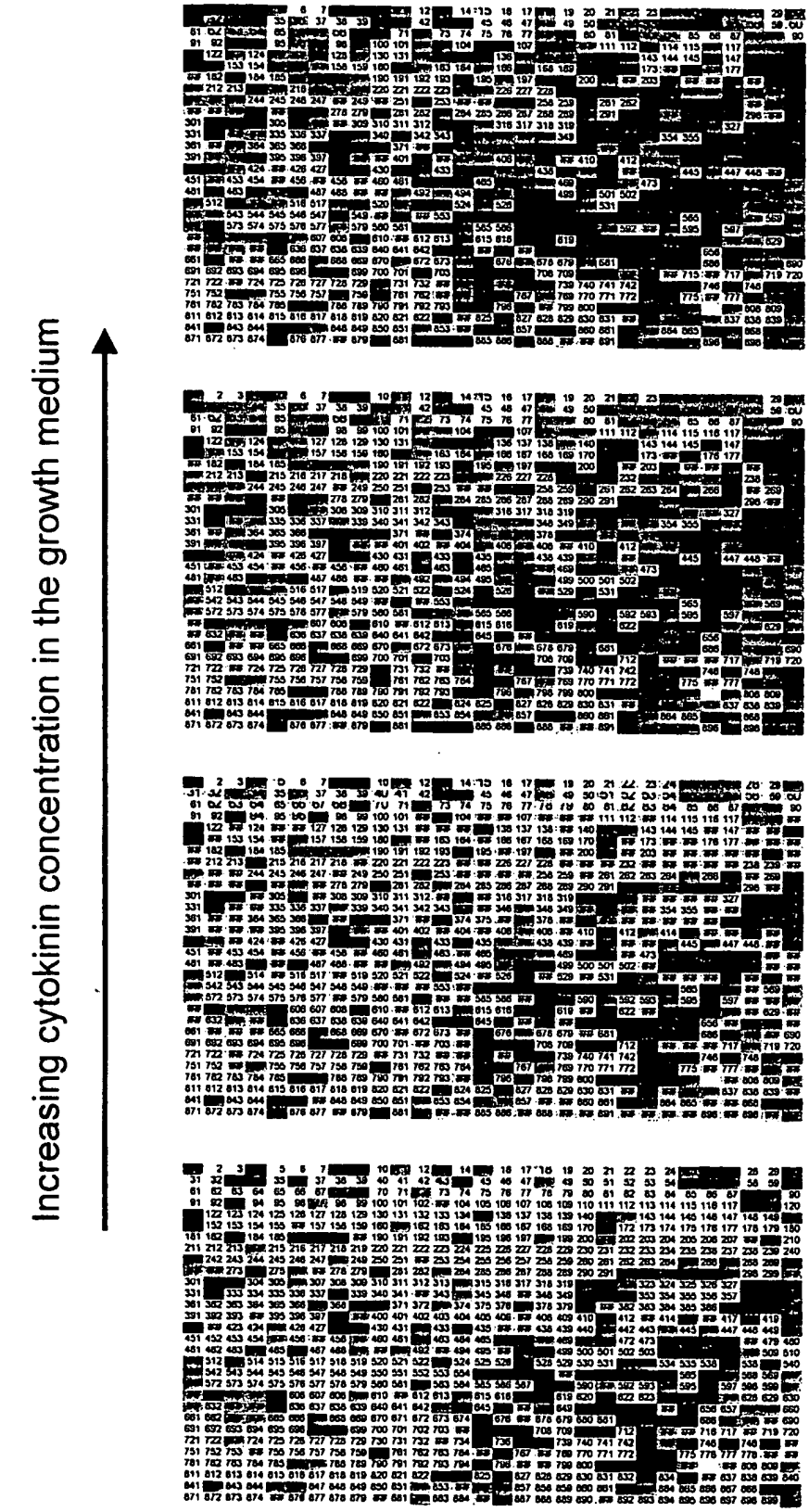
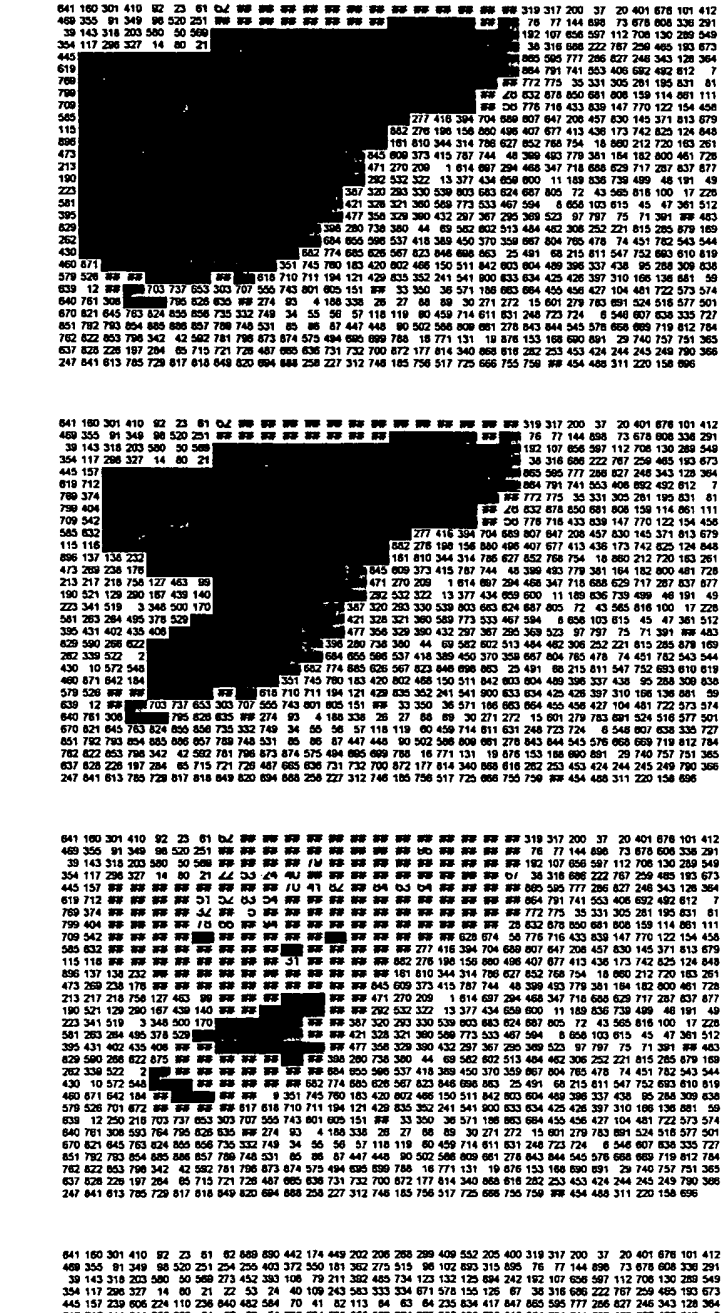

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium
Microarray data before rearrangement.

FIG. 2D



Increasing cytokinin concentration in the growth medium

Increasing auxin concentration in the growth medium

Microarray data after rearrangement

FIG. 3A

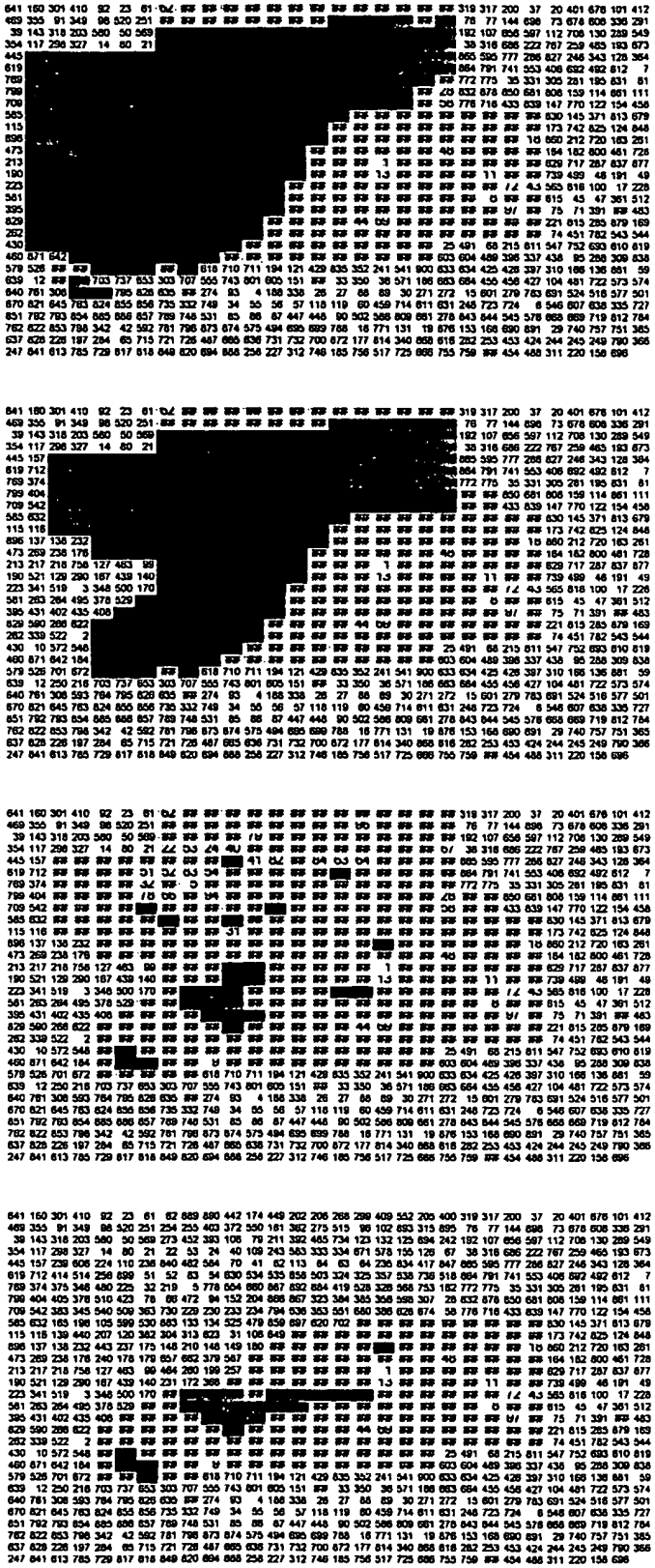
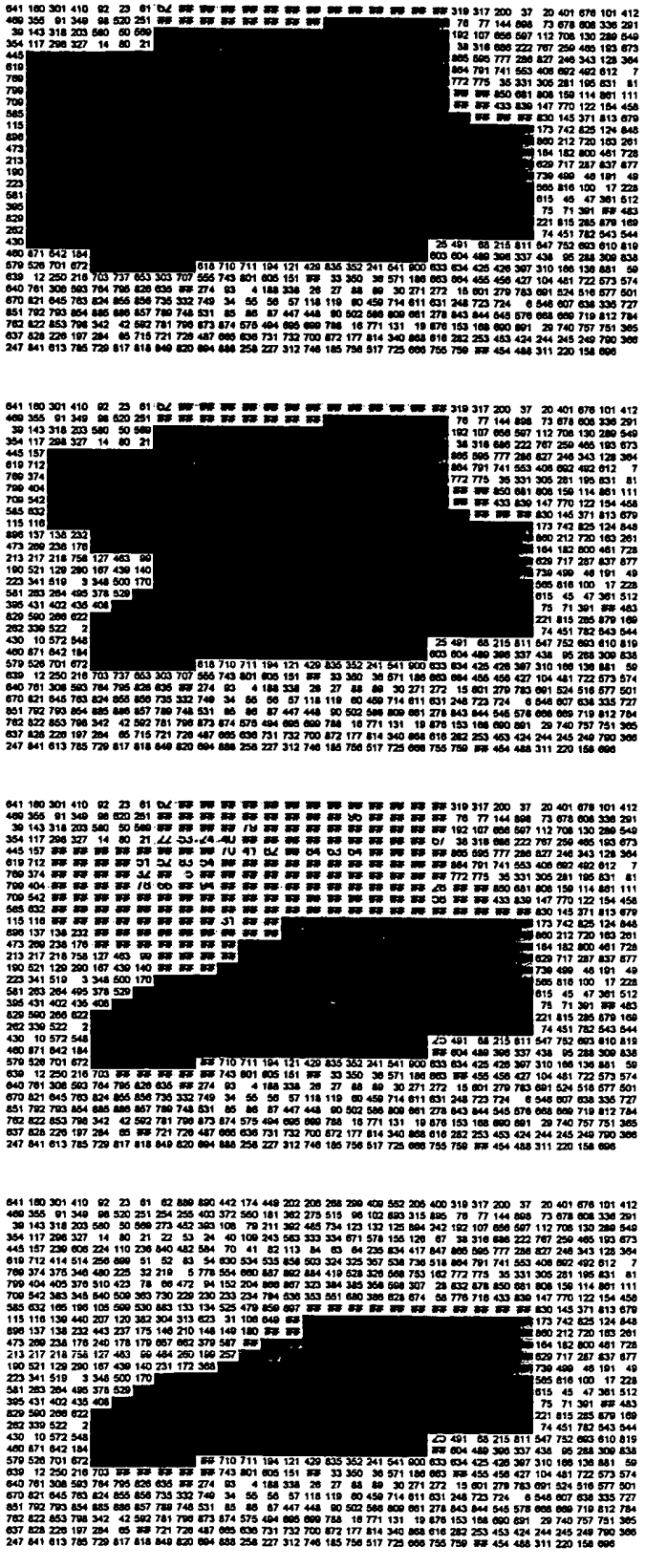

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium
Microarray data after rearrangement

FIG. 3B



Increasing cytokinin concentration in the growth medium

Increasing auxin concentration in the growth medium

Microarray data after rearrangement

FIG. 3C

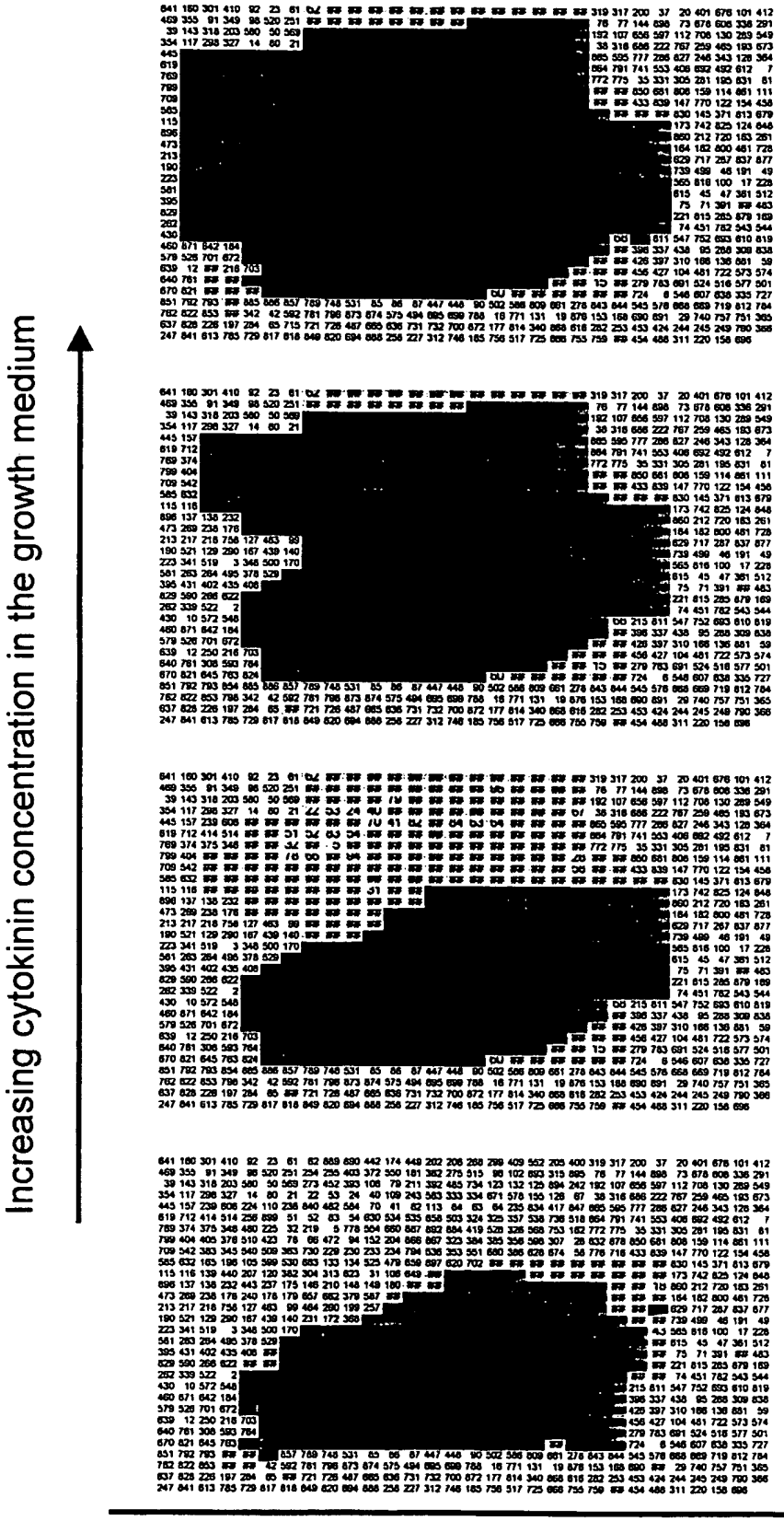

Increasing cytokinin concentration in the growth medium
Microarray data after rearrangement
Increasing auxin concentration in the growth medium

FIG. 3D

The rearrangement indicates the function of the different genes

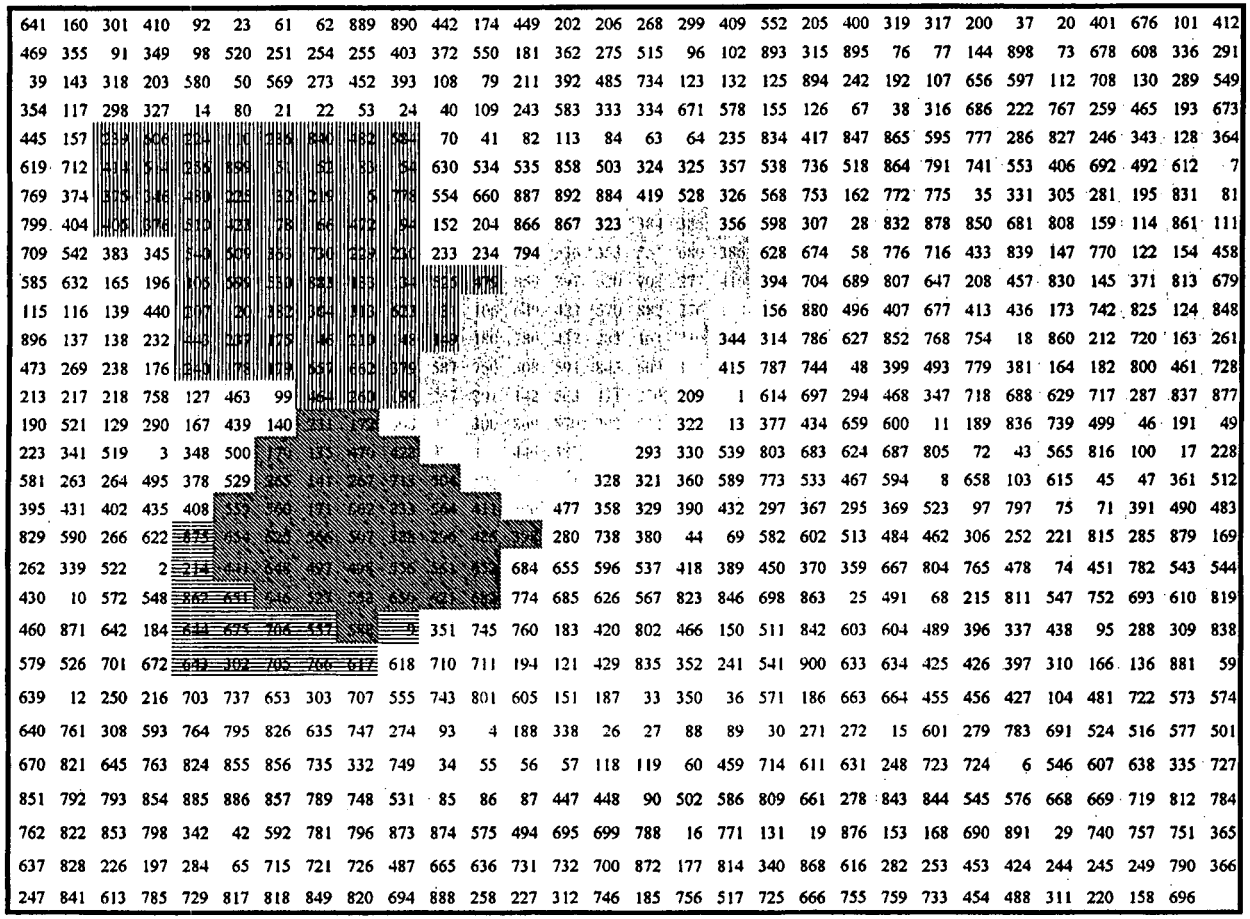


FIG. 4

genes involved in:

- Cell division
- general
- roots
- Vascular tissue
- shoots
- White callus
- Chloroplasts
- Green callus
- Non of the above

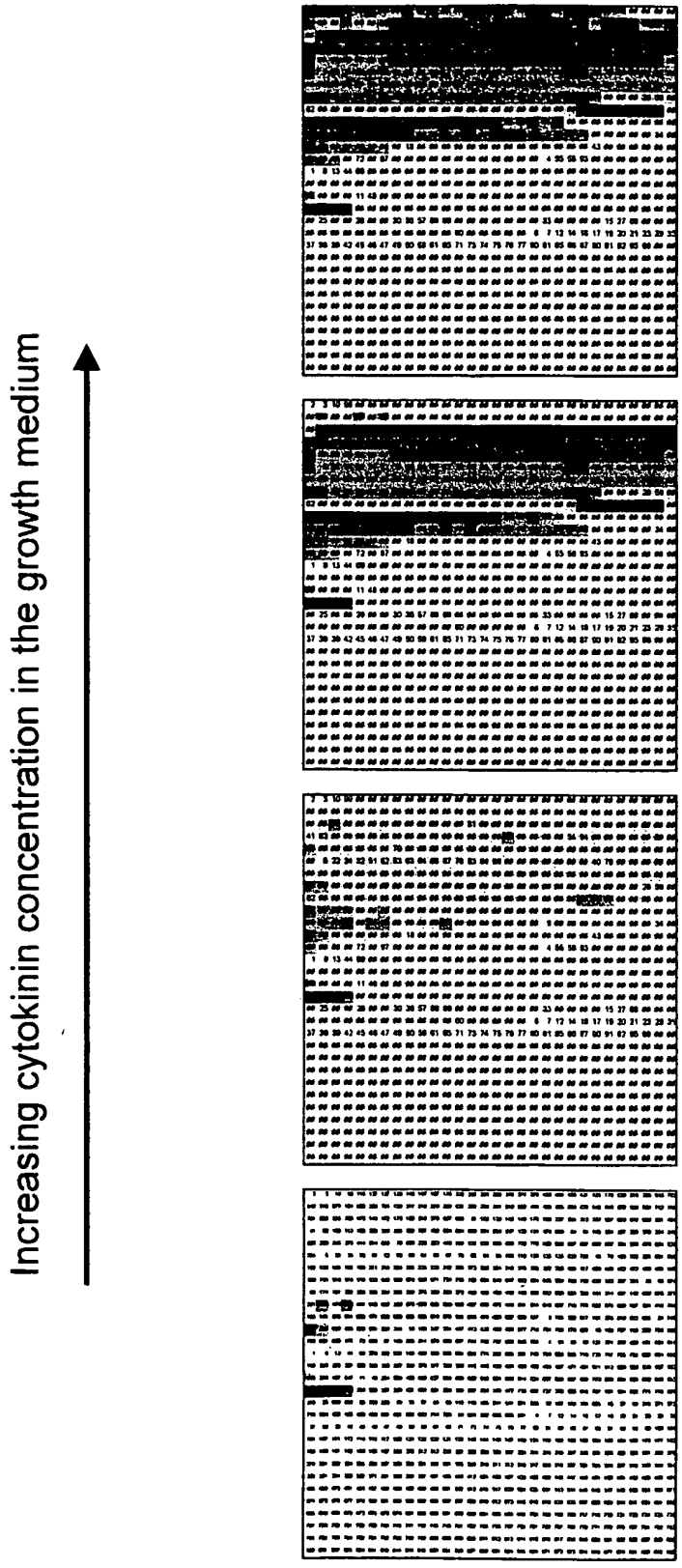

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium

FIG. 5A

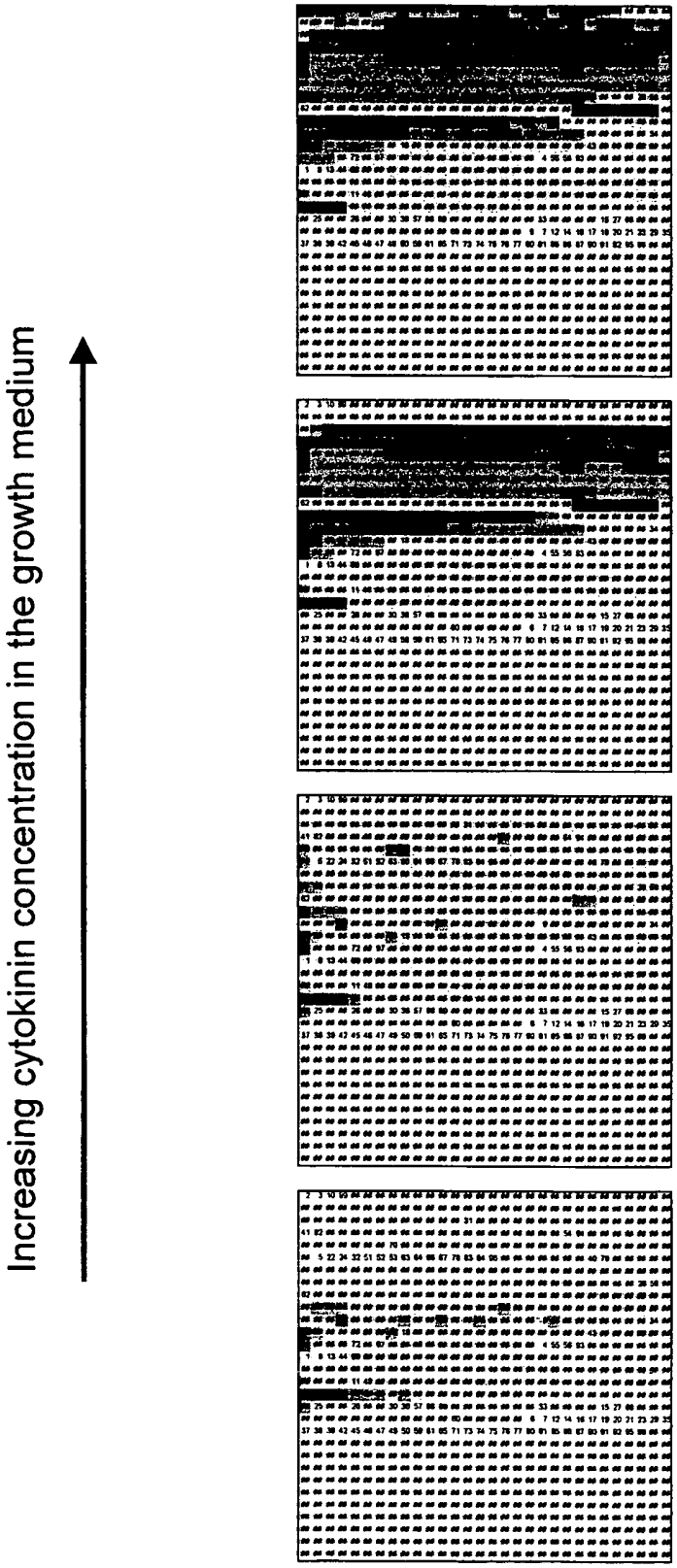

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium

FIG. 5B

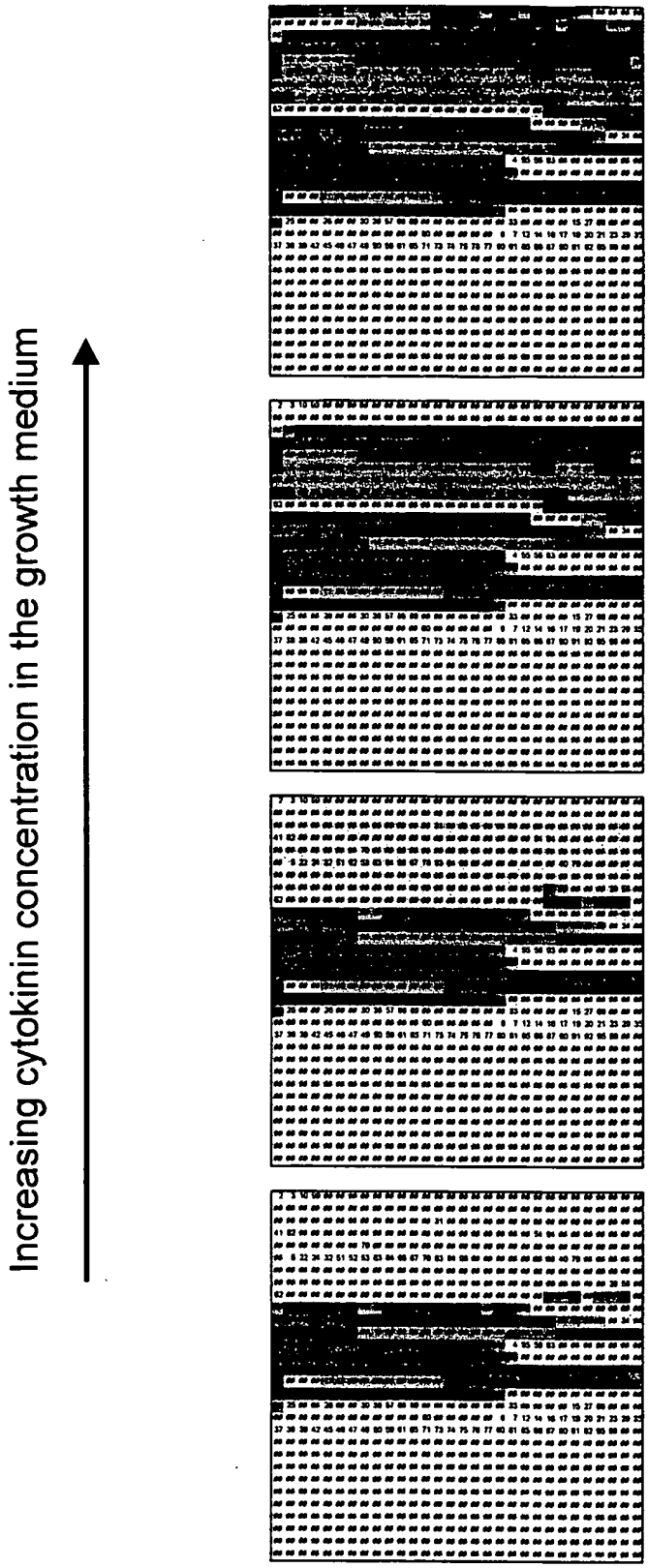

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium

FIG. 5C

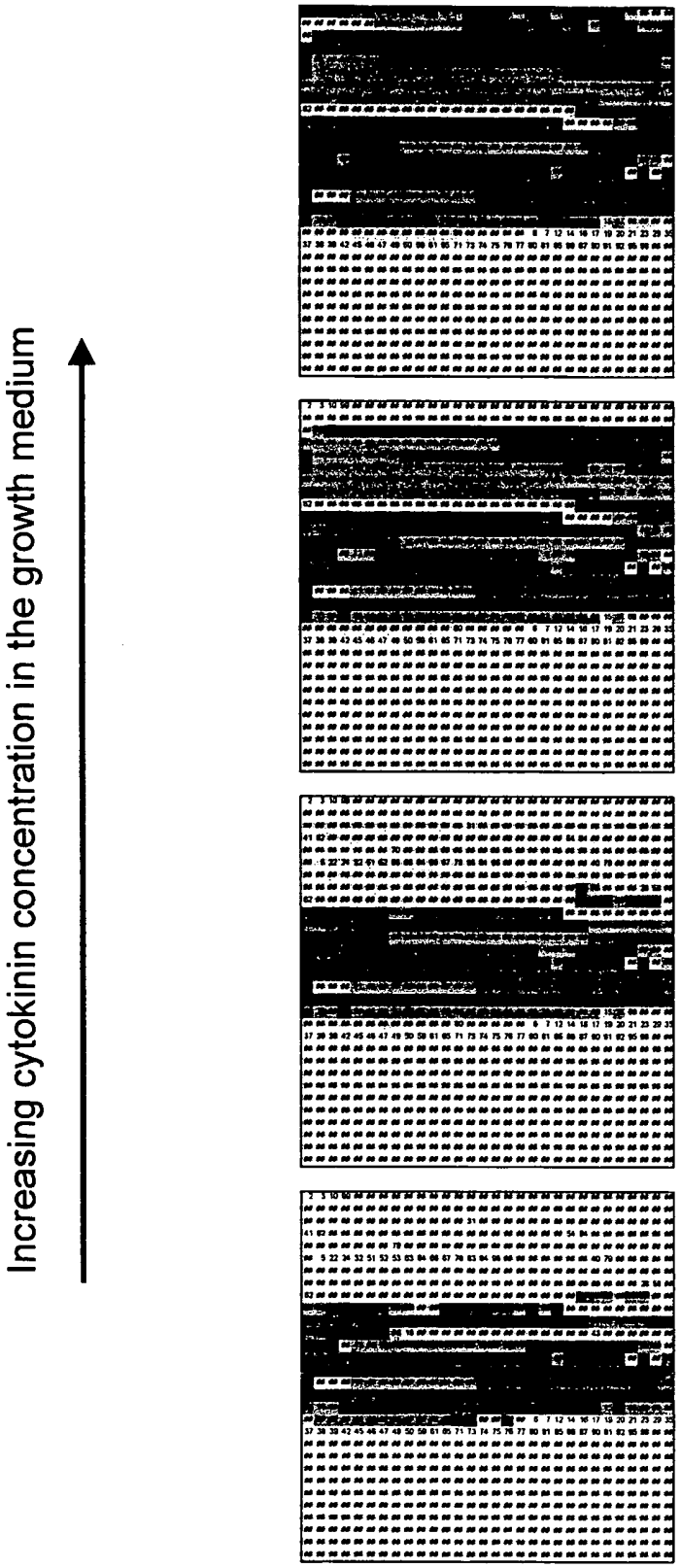

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium

FIG. 5D

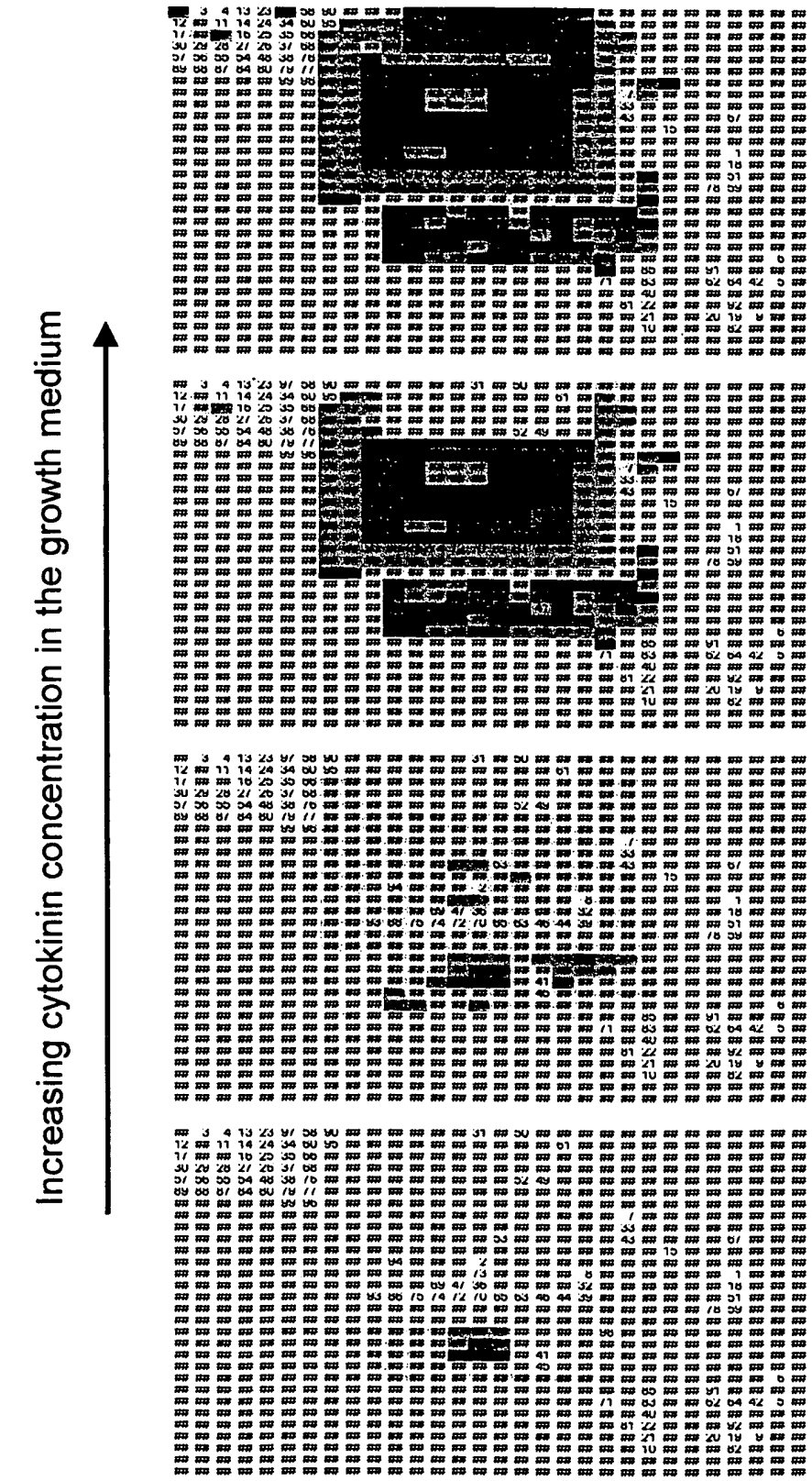
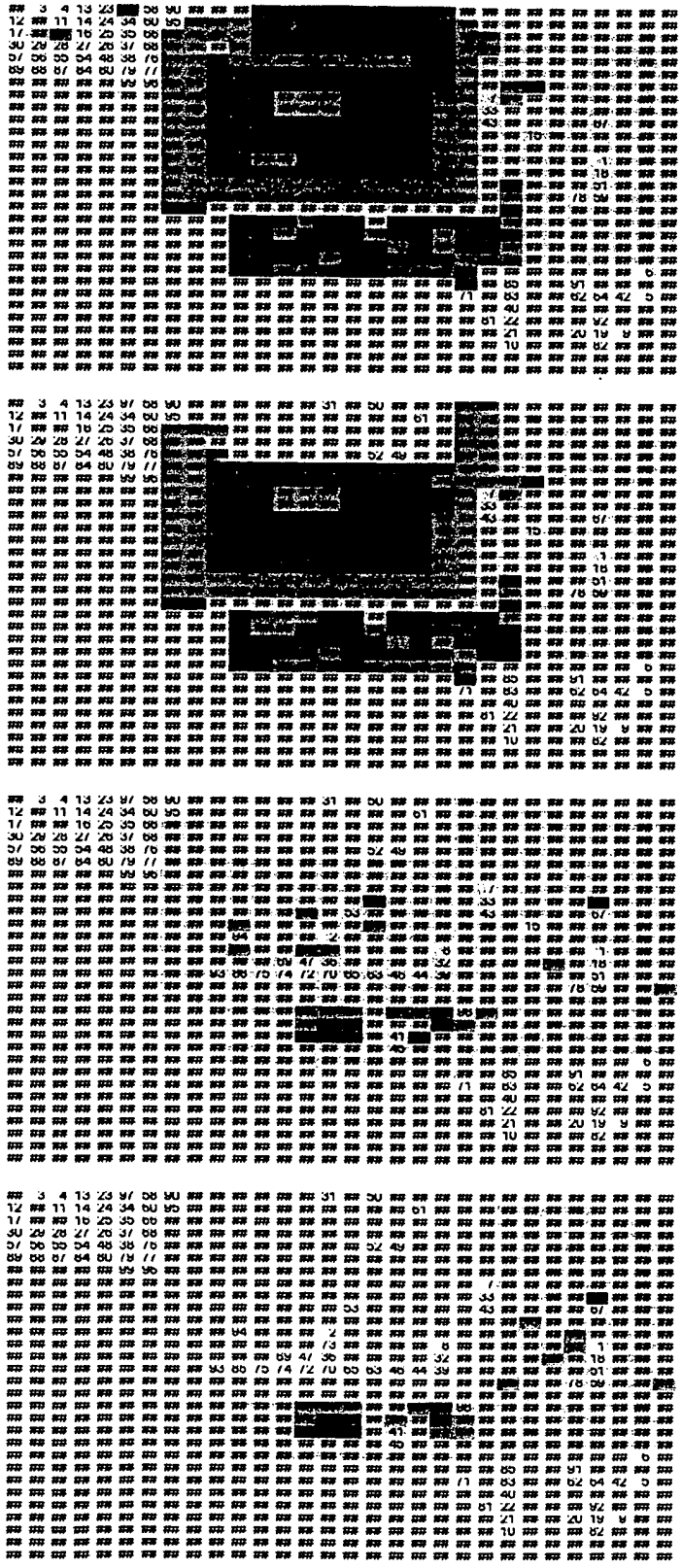

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium

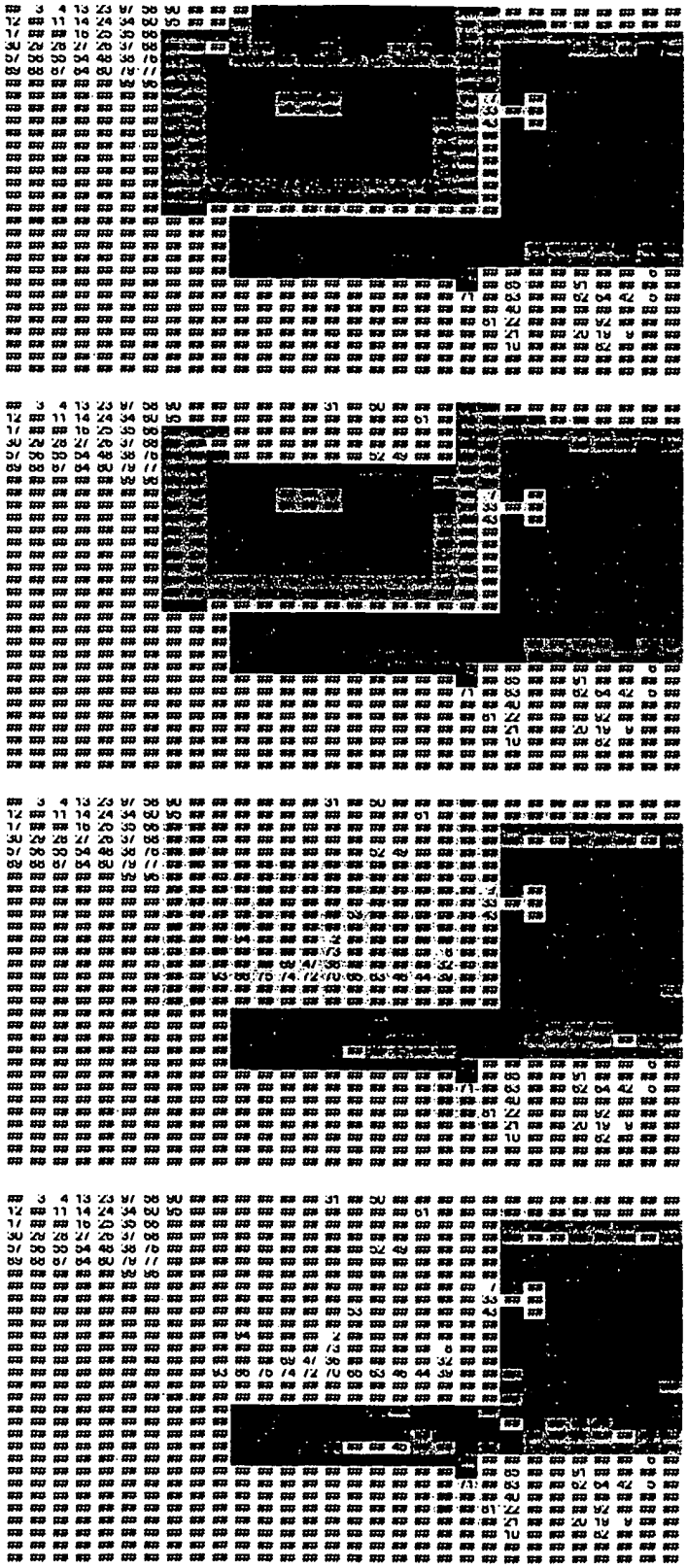
FIG. 6A



Increasing cytokinin concentration in the growth medium

Increasing auxin concentration in the growth medium

FIG. 6B



Increasing cytokinin concentration in the growth medium

Increasing auxin concentration in the growth medium

FIG. 6C

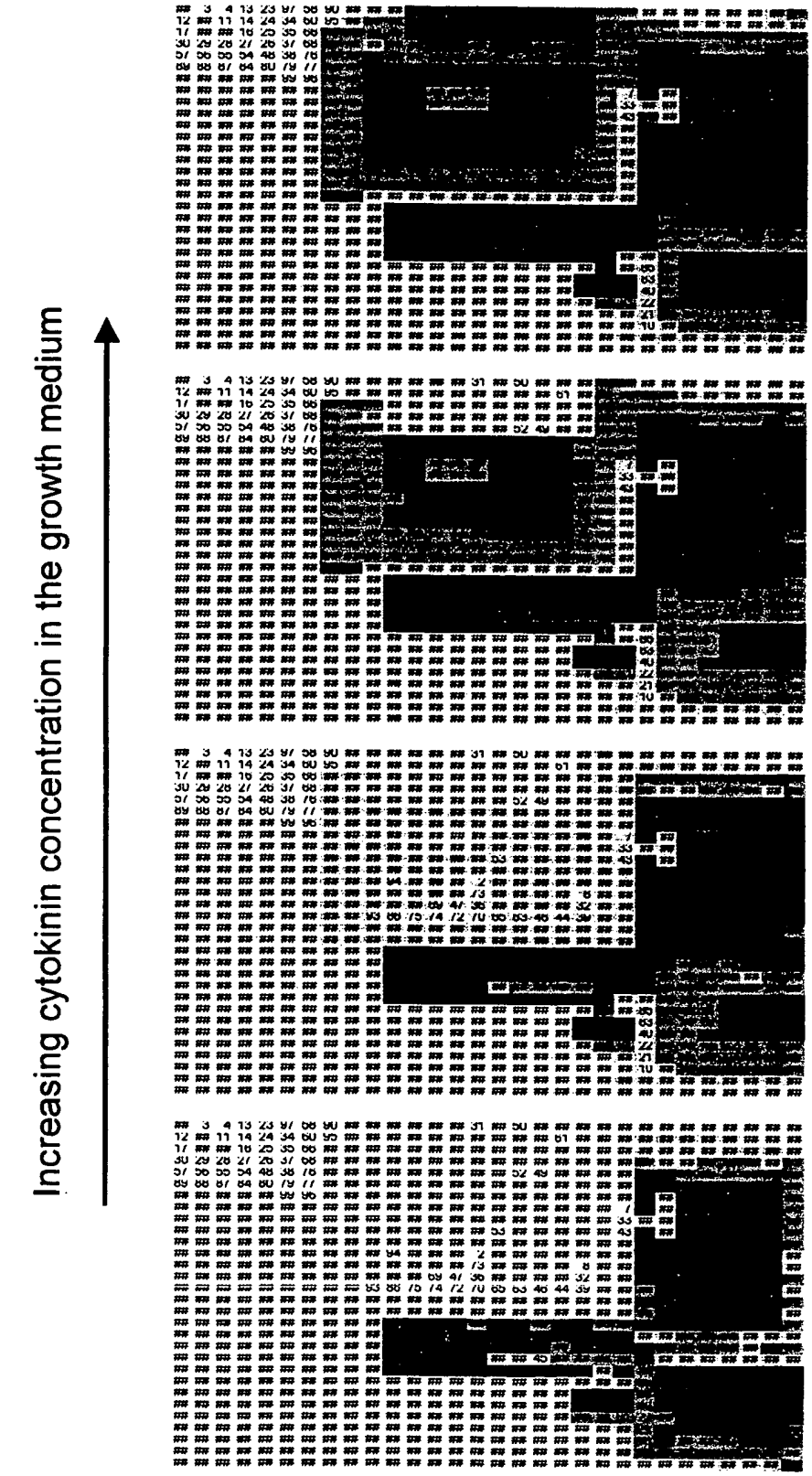

Increasing cytokinin concentration in the growth medium
Increasing auxin concentration in the growth medium

FIG. 6D